Project 4: Bank Marketing

02/28/2020

1 Introduction

1.1 Background

Marketing selling campaigns constitute a typical strategy to enhance business. Companies use direct marketing when targeting segments of customers by contacting them to meet a specific goal. Telephone (fixed-line or mobile) is one of the most widely used. Technology enables rethinking marketing by focusing on maximizing customer lifetime value through the evaluation of available information and customer metrics, thus allowing us to build longer and tighter relations in alignment with business demand [1].

The dataset is downloaded from UCI Machine Learning Repository and is related to direct marketing campaigns of a Portuguese Banking institution [2]. These campaigns were based on phone calls. Often, more than one calls were done to the same client to access if their product "term deposit" will be subscribed (yes) or not subscribed (no). There were 4 datasets in it from which bank-additional-full.csv is used that has all examples (41188) and 20 inputs ordered by date (from May 2008 to November 2010). There are 20 input variables and 1 output variable (desired target). These dataset attributes denote customer data, socio-economic data, telemarketing data and some other data. Some attributes are numerical, and some are categorical. The dataset was loaded in R Studio and checked for any missing values using is.na function and found that it didn't have any missing values, so we have a clean dataset.

1.2 Statistical questions of interest

Our main goal is to build prediction models based on the data set to predict response variables, which a customer would subscribe to a bank long-term deposit or not. Based on all 20 variables from the dataset, we divide the data into training data and test data. After that, we select variables by stepwise method to build models. We will fit two models through training data, including the logistics regression model and the random forest, respectively. Then, we will test the models on test data. Our interest in this project is to calculate the prediction accuracy of these two models and compare the gap performance of the two models.

2 Analysis Plan

2.1 Descriptive Analysis

Table.1 Features description of the Bank Marketing Dataset (BMD).

Feature	Description	Attribution
`y`	Desired target. Has the client subscribed a term deposit?	binary
`age`	Client Age	numeric
`job`	Type of Job	categorical
`marital`	Client's marital status	categorical
`education`	Client's education	categorical

Feature	Description	Attribution
default	Has credit in default?	categorical
housing	Has housing loan?	categorical
loan	Has personal loan?	categorical
contact	Last contact month of year	categorical
month	Month of last contact with client	categorical
day_of_week	Last contact day of the week	categorical
duration	last contact duration, in seconds	numeric
campaign	Number of contacts performed during this campaign and for this client	categorical
pdays	Number of days that passed by after the client was last contacted from a previous campaign	numeric
previous	Number of client contacts performed before this campaign	numeric
poutcome	Outcome of the previous marketing campaign	categorical
emp.var.rate	Quarterly employment variation rate	numeric
cons.price.idx	Monthly consumer price index	numeric
cons.conf.idx	Monthly consumer confidence index	numeric
euribor3m	Daily euribor 3-month rate	numeric
nr.employed	Quarterly number of employees	numeric

By description, `duration` should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model, so this attribute highly affects the output target. More than 95% values in `pdays` are 999, which means client was not previously contacted. Only 3 cases are `yes` in `default`, and more than 20% are unknown. Thus, we remove these three variables.

2.2 Predictive Model

To predict, given the seventeen features, if the bank term deposit would be or not subscribed, we used logistic regression and random forest. As mentioned before, the BMD consists of 41188 observations. A random sample of 10% of this size, 4119 observations, was withdrawn to be used as a test dataset. The rest, 37069 observations, was used as a train dataset.

2.2.1 Logistic Regression

The main goal of our project is to predict whether the client has subscribed a term deposit. In other words, the response variable has a binary outcome. Therefore, we decide to choose the logistic model to fit the dataset. The model is given as follows:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

where $p(x)$ is the probability that the dependent variable equals a case, given some linear combination of the predictors; $x_i, i = 1, \cdots, n$ is the predict variables; β_0 is the intercept from the linear regression equation and $\beta_i, i = 1, \cdots, n$ is the regression coefficient.

Model assumptions: (1) The outcome is a binary or dichotomous variable like yes or no, positive or negative, 1 or 0; (2) There is no influential value (extreme values or outliers) in the continuous predictors; (3) There is no high intercorrelations (i.e. multicollinearity) among the predictors.

To test the significance of the features we use the Akaike information criterion (AIC). Given a collection of models, the AIC estimates the quality of each model, relative to each of the other models. Therefore, AIC provides a mean for model selection.

2.2.2 Model Diagnostic

According to the assumptions of the logistic model, we will first check if the outcome is binary. Influential values are extreme individual data points that can alter the quality of the logistic regression model. Therefore, Cook's Distance plot and Studentized Residual plot will be figured to find whether there are influential values and outliers in the dataset. Finally, Multicollinearity will be considered by VIF values becuase it corresponds to a situation where the data contain highly correlated predictor variables.

2.2.3 Random Forest

Random forests are an ensemble learning method that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

Random forests differ in only one way from tree bagging: they use a modified tree learning algorithm that selects, at each candidate split in the learning process, a random subset of the features. Tree bagging repeatedly selects a random sample with replacement of the training set and fits trees to these samples. This bootstrapping procedure leads to better model performance because it decreases the variance of the model, without increasing the bias. In R, the main implementation of random forest is found in the randomForest library [3].

2.3 Comparison

The main measure that can be used to compare different algorithms is the Receiver Operating Characteristic curve, i.e. ROC curve. A graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the specificity, true negative rate, against the sensitivity, true positive rate, at various threshold settings. In R, the main implementation of the ROC curve is found in the pROC library [4].

When dealing with ROC curves the main measure returned is the Area Under the Curve (AUC), that is qual to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

3 Results

3.1 Descriptive Analysis

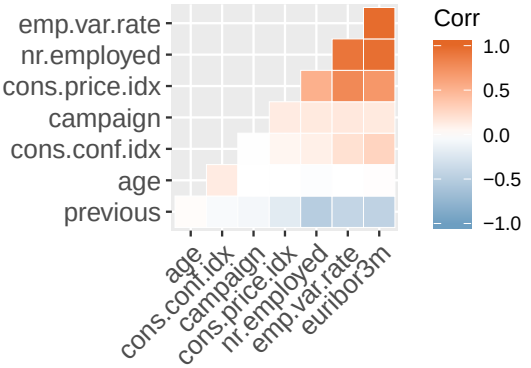

Figure.1 correlation matrix for all the quantitative features presented in the Bank Marketing Dataset (BMD).

In Figure 1 we see the two-by-two correlations for all the eight numerical features in the BMD. There are high correlation among `emp.var.rate`, `euribor3m` and `nr.employed`. To avoid high collinearlity, we only remain `nr.employed` for analysis.

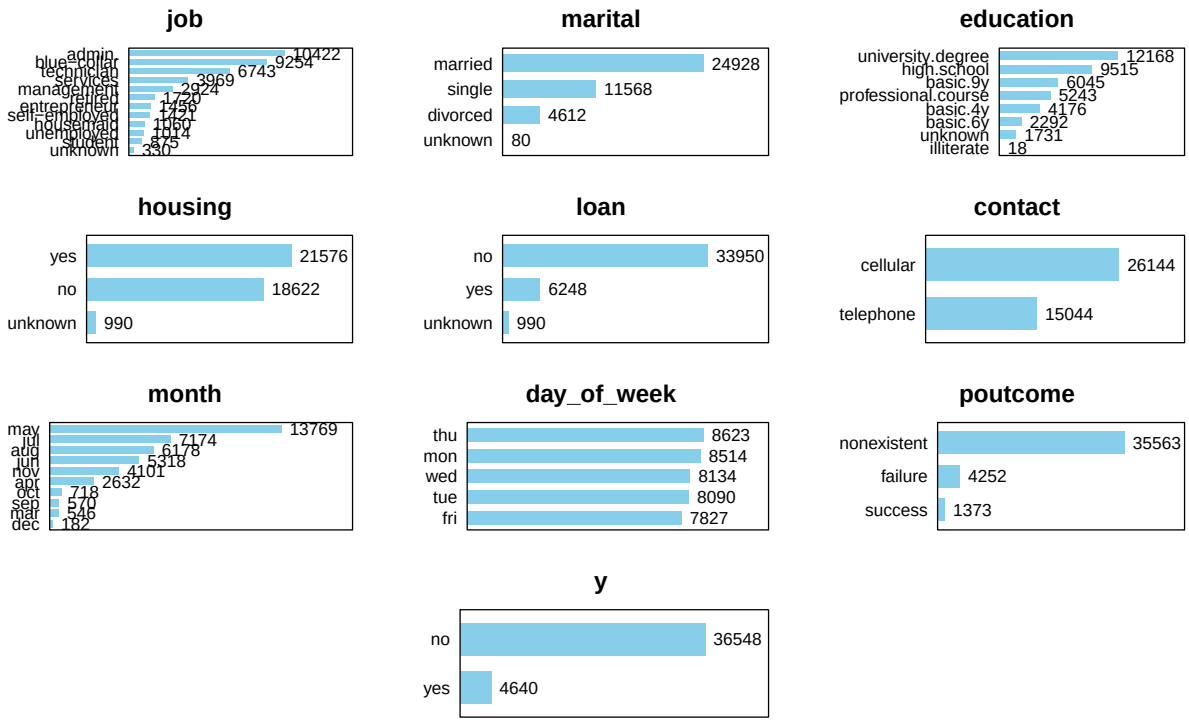

Figure.2 Bar plots for all the qualitative features presented in the Bank Marketing Dataset (BMD).

Already in Figure 2 we have the frequencies for each level of the categorical features in the BMD. First, we see that the desired target is unbalanced, with more than 85% of the observations corresponding to clients that didn't subscribed to a term deposit. An equilibrium between levels is only present in the `day.of.week` last contact feature. By this Figure we can also see that the last contact of most of the clients was in may (`month` feature), that most of the clients have a nonexistent previous marketing campaign (`poutcome` feature), that they are married (`marital` feature) and that most have a `job` in the administrative sector.

3.2 Logistic Regression

From fifteen features, logistic regression model finished with nine: `job`, `contact`, `month`, `day_of_week`, `campaign`, `poutcome`, `emp.var.rate`, `cons.price.idx`, `cons.conf.idx` and `nr.employed`. Keeping a variable means that the feature is statistically significant, in describing the difference between the classes of the desired target - if the bank term deposit would be or not subscribed. The fitted value on test data and the performance will be discussed together with the result from random forest in section 3.4.

3.3 Model Diagnostics

Our desired target `y` means that whether the client has subscribed a term deposit or not. Therefore, the outcome satisfies the binary outcome assumption.

Influential values are extreme individual data points that can alter the quality of the logistic regression model. The Cook's distances are much smaller than 1, meaning that there is not influential case in the dataset. To check whether the data contains potential outliers, the standardized residual error can be inspected. If an observation has an externally studentized residual that is larger than 3 (in absolute value), we can call it an outlier. Therefore, from the plot, we can verify that there is none outliers in the dataset [5].

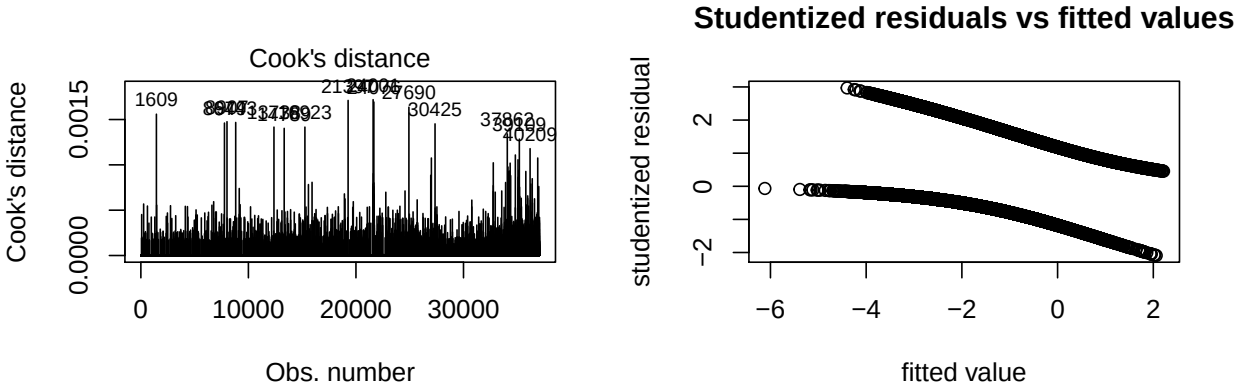

Figure.3 Cook's Distance and Studentized residuals vs fitted values plot

Multicollinearity is an important issue in regression analysis and should be fixed by removing the concerned variables. As a rule of thumb, a VIF value that exceeds 10 indicates a problematic amount of collinearity. In our model, only the variable `emp.var.rate` has a VIF value a little larger than 10. However, it will not influence the accuracy of prediection, so we keep it in model.

Table.2 VIF value for logistic model

Fit	job	contact	month	week	campaign	poutcome	emp.var.rate	cons.price.idx	cons.conf.idx	nr.employed
VIF	1.01	1.54	1.20	1.01	1.02	1.07	11.99	7.31	1.61	8.67

3.4 Comparison between Logistic Regression and Random Forest

Table 3 is the confusion table between fited value and true value for test data in logistic regression and random forest model. Table 4 shows accuracy, specificity and sensitivity in different models. The logistic regression model has a higher sensitivity, and the random forest model has a slightly higher specificity. Moreover, the accuracy of them are nearly the same. Both of them have a good specificity (i.e. true negative rate), but a not good sensitivity (i.e. true positive rate), which means that these classifiers rarely register "yes" for people who have no intention to a term deposit. Moreover, they are more likely to overlook people who are willing to subscribe to a term deposit and predict that they do not subscribe it. The AUC for each model is presented in Figure 4. The highest is obtained with the logistic regression model. From the above, it shows that the logistic regression model has a better performance than the random forest model.

Table.3 Confusion table between prediction and true value in logistic regression and rondom forest

logistic regression		True	True	random forest		True	True
		yes	no			yes	no
Predicted	yes	99	321	Predicted	yes	134	286
Predicted	no	71	3628	Predicted	no	126	3573

Table.4 Specificity, sensitivity and accuracy for each model in the test Bank Marketing Dataset.

Model	Accuracy	Specificity	Sensitivity
Logistic regression	**0.9048**	0.9187	**0.5824**
Random forest	0.8992	**0.9259**	0.5154

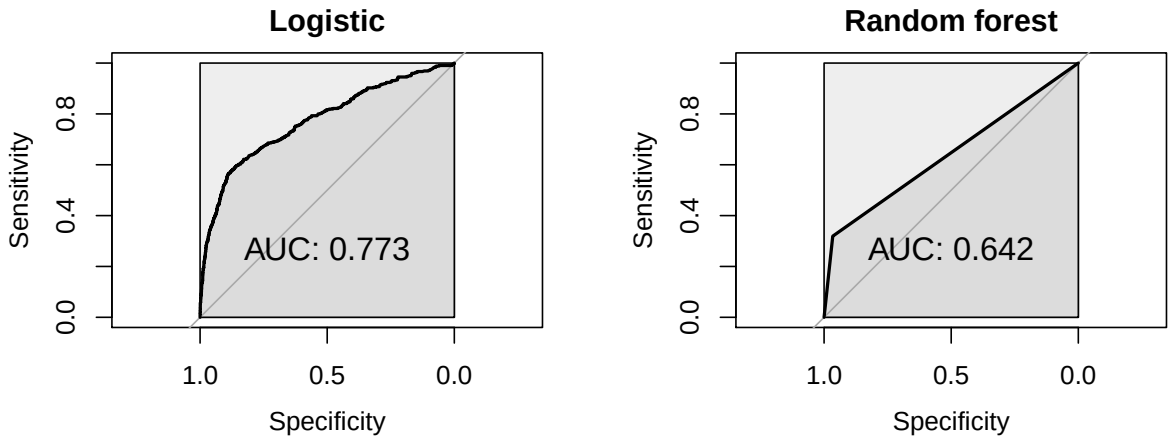

Figure.4 ROC curve for each model (in the test) with respective AUC and thresholds.

4 Discussion

Looking from the above, we find that the performance of two models in accuracy and specificity is nearly consistent. However, the sensitivity of the logistics model is higher than of the random forest model. From literature [6], it found that when increasing the variance in the explanatory and noise variables, logistic regression consistently performed with higher overall accuracy compared to random forest. The logistic regression performs better when the number of noise variables is less than or equal to the number of explanatory variables, and it has a higher true positive rate. In our case, the VIF value of all variables is less than 10 except `emp.var.rate` whose value is slightly greater than 10. Moreover, the proportion of noise variables, including campaign, p-days, previous and poutcome is relatively small. Therefore, it explains the consistent performance in accuracy and specificity and the gap in sensitivity.

It should be noticed that our dataset is unbalanced, and over 85% observations where value of y is `no`. Under the situation of unbalanced data, the predictive results are not accurate. Also, the sensitivity is not good in unbalanced data [6]. To solve this issue, we can undersample the majority class method to deal with it in further investigation.

5 Appendix. Reference

[1] [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

[2] Bank Marketing Data Set. UCI machine learning repository. https://archive.ics.uci.edu/ml/datasets/Bank+Marketing.

[3] Liaw, A. & Wiener M. (2002). Classification and Regression by randomForest. R News 2(3), 18–22. http://CRAN.R-project.org/doc/Rnews/.

[4] Robin, X., Turck, N., Hainard, etc. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics, 12, p. 77. DOI: 10.1186/1471-2105-12-77. http://www.biomedcentral.com/1471-2105/12/77/.

[5] Measures of Influence. https://cran.r-project.org/web/packages/olsrr/vignettes/influence_measures.html.

[6] Kirasich, Kaitlin; Smith, Trace; and Sadler, Bivin (2018) "Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets," SMU Data Science Review: Vol. 1 : No. 3 , Article 9. Available at: https://scholar.smu.edu/datasciencereview/vol1/iss3/9.